Machine Learning – In the Fashion Industry

Group 6

OUR TEAM

Maria

Aina

Levan

Subiksha

TABLE OF CONTENTS

01 → **Introduction**

02 → **Data gathering**

03 → **Data cleaning & Exploration**

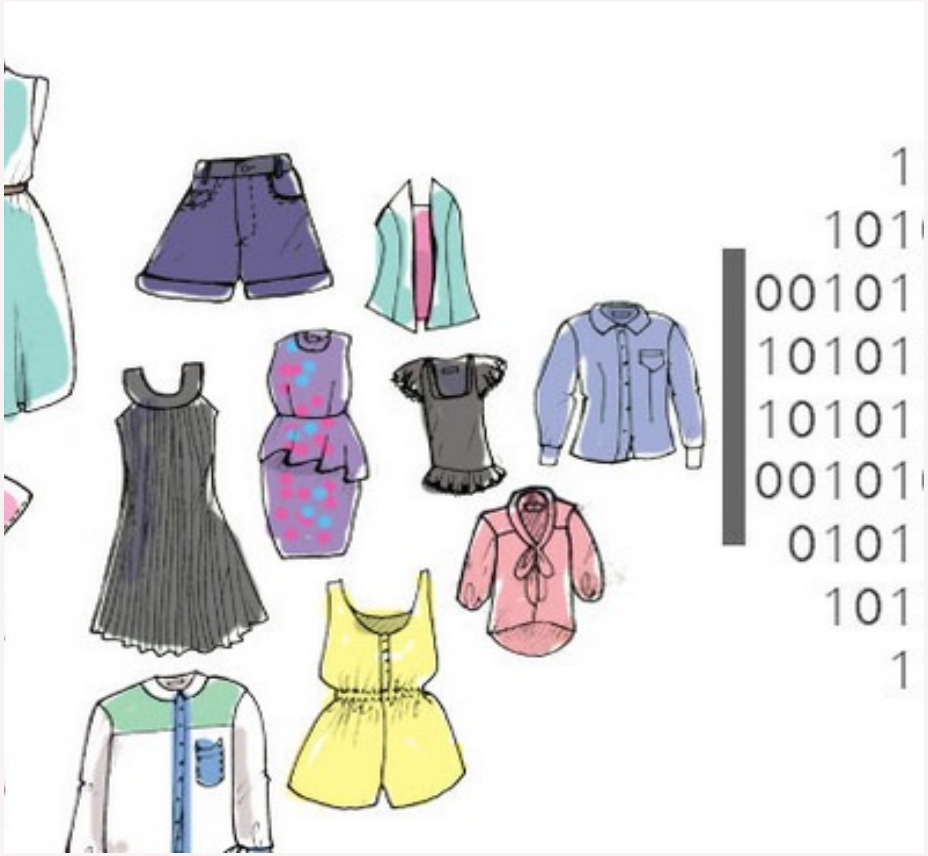
04 → **Optimization**

05 → **Visualization**

06 → **Conclusion**

Data Gathering

- Kaggle dataset: Fashion Dataset UK-US
- 20 columns: Price
  Season
  Brand
  Style
  Category
  Number of reviews
- Over 1 millions rows



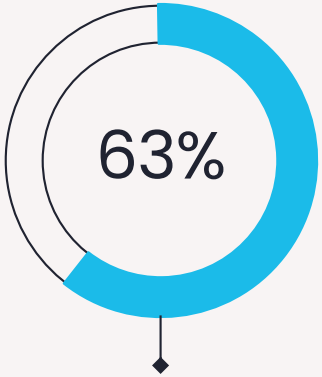
The Goal:

As a fashion wholesaler, what category and size of clothes to buy for the all seasons?

Data Cleaning and Exploration

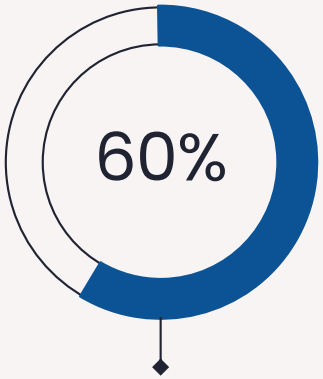
- Drop non-value-added column
- Drop N/A values
- Normalize rating customer columns
- Get dummies function for style attributes
- Data type
- Data description (count, mean, min, max)
- Correlation Matrix
- Boxplot for outliers (matplot lib)

The testing phase



Decision Tree

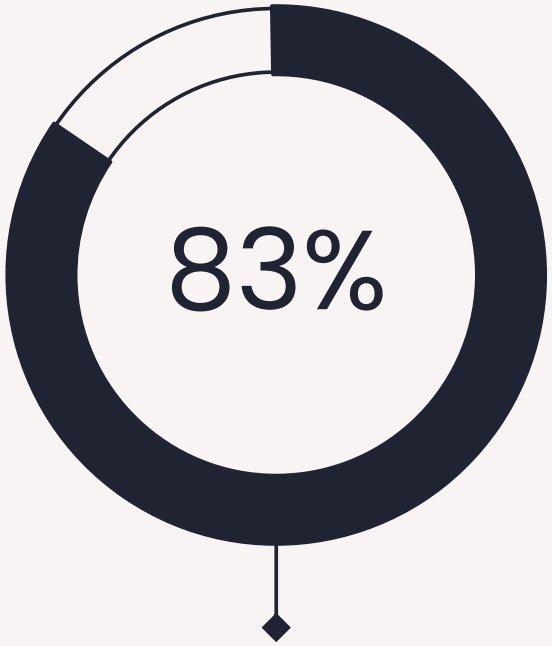
Was affected by our unbalanced dataset



XGBoost

Random Forest gave a better accuracy score

Random Forest



Definition:

- Classification algorithm
- Ensemble type of decision tree

Library used:

- Pandas
- train_test_split
- sklearn.metrics import classification_report
- RandomForestClassifier
- cross_val_score


```
Cross-Validation Scores: [0.8337125  0.8336625  0.8319875  0.83308125 0.83288125]
Test Set Score: 0.8329
              precision    recall  f1-score   support

           0       0.87      0.88      0.88    133547
           1       0.76      0.73      0.74     66453

    accuracy                           0.83    200000
   macro avg       0.81      0.81      0.81    200000
weighted avg       0.83      0.83      0.83    200000
```

Why Random Forest?

Argument 1

Large number of observations

Argument 2

Data mix – Numerical and categorical

Argument 3

Highest performance from our testing phase

Argument 4

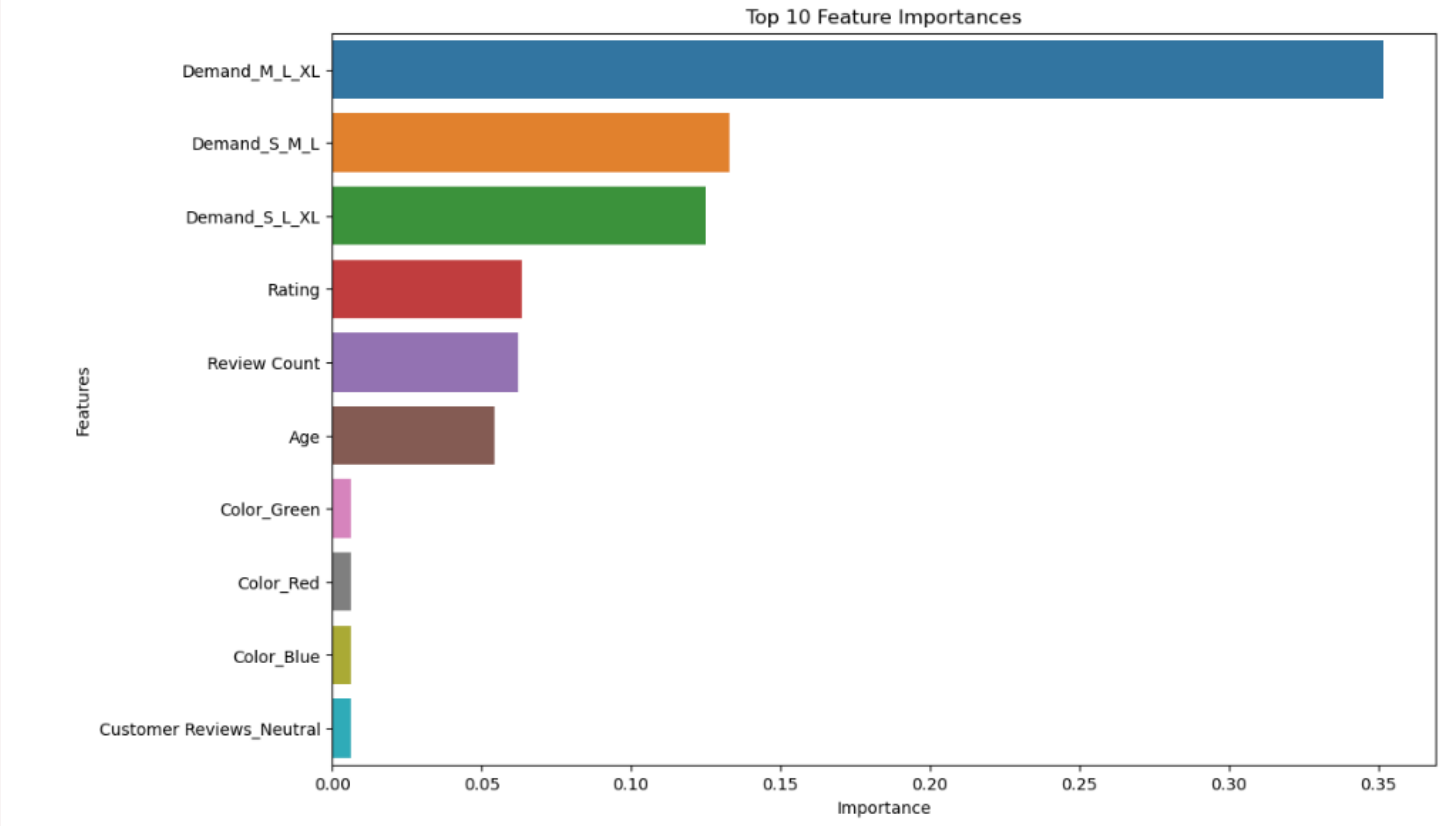
Builds trees sequentially, each correcting the errors of the previous one

Optimization

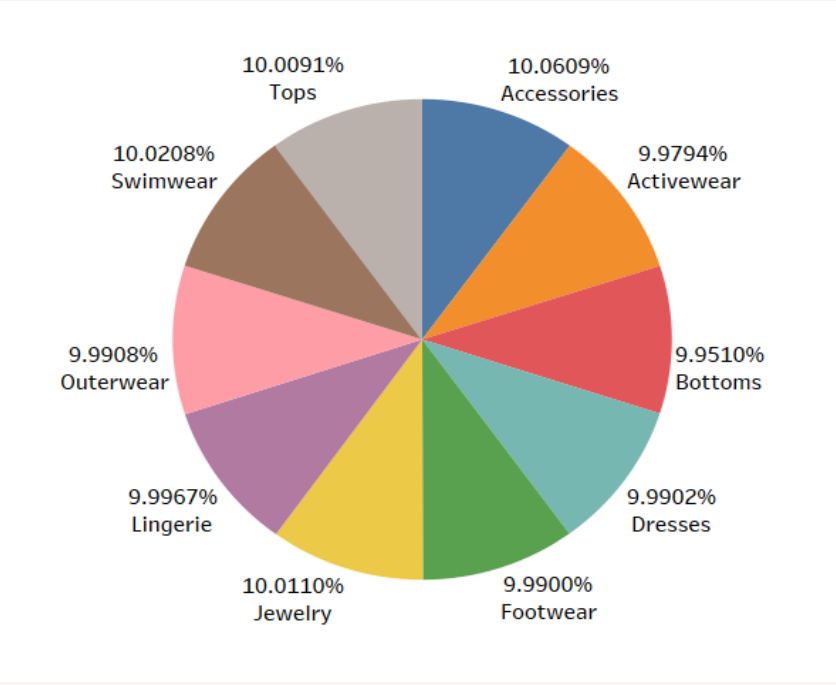
Steps followed

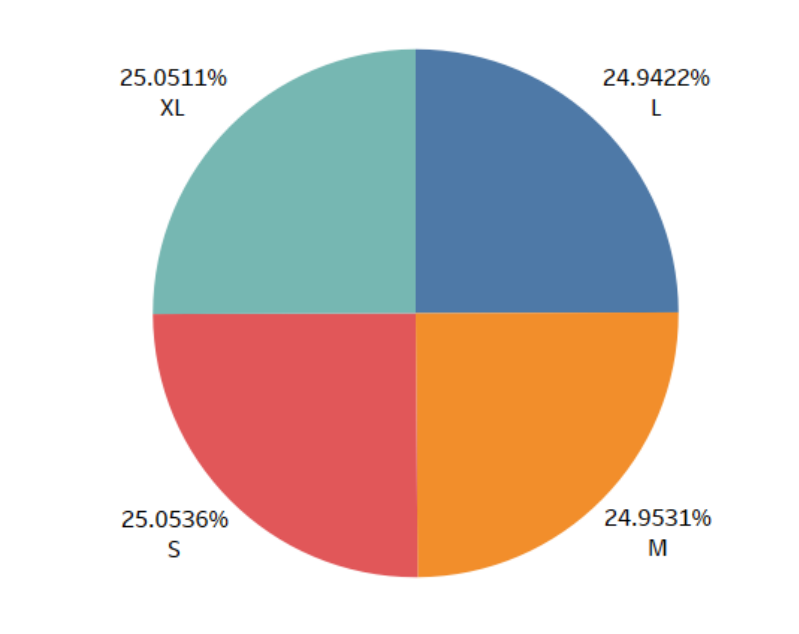
- Unbalanced Data – Binary version
- SMOTE - Synthetic Minority Over-sampling Technique
- Small improvement

Top 10 Features Importance



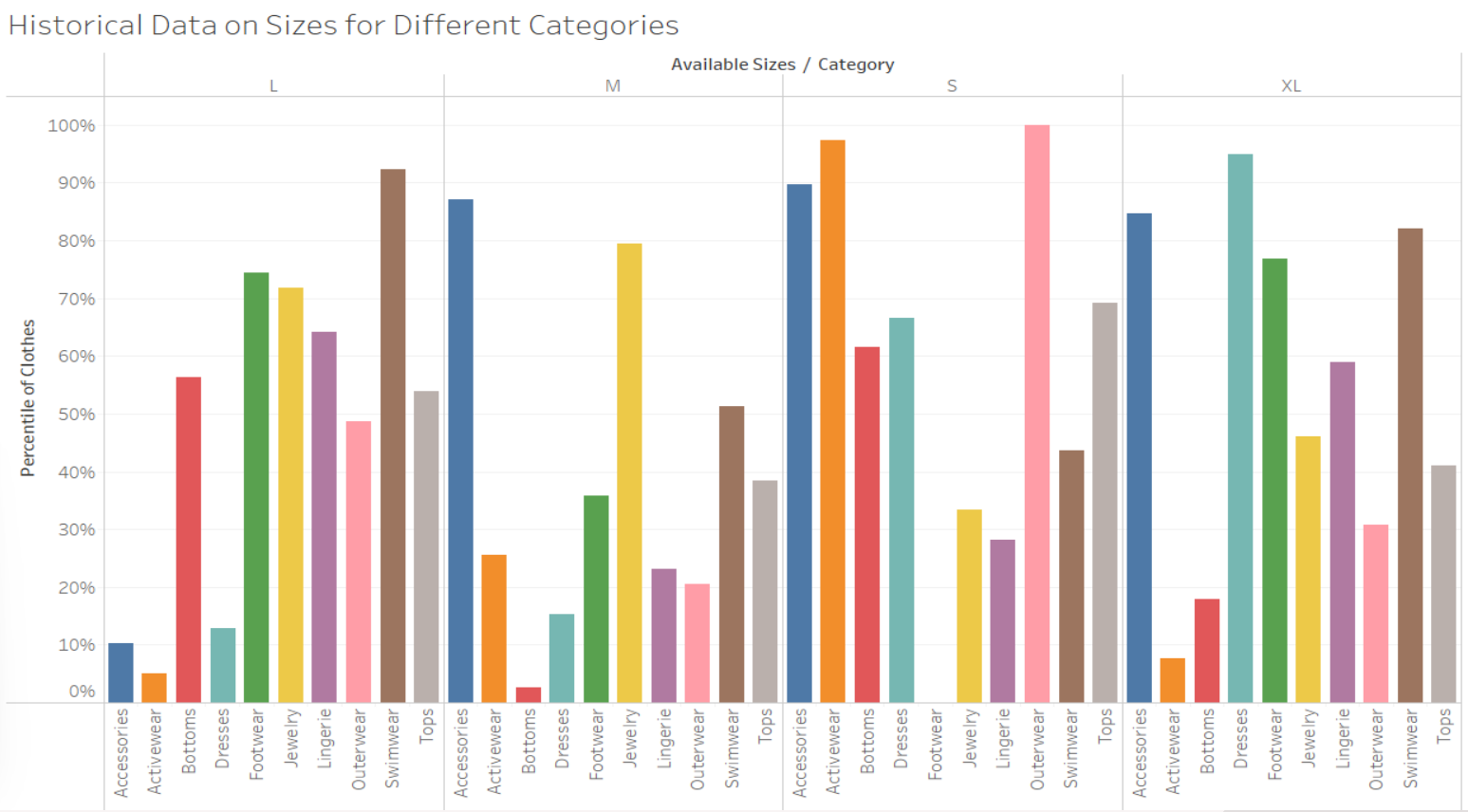
Categories of Clothes





~ 10 categories
~ Equal size distribution

Demand of Sizes for Categories



THANKS!

Question?